SGI Cray Origin2000 Architecture

Overview

- First product from the merger of SGI & Cray
- Successor to the PowerChallenge Symmetric Multi-Processor, SMP, system
- Origin 2000 is S2MP, Scaled Shared-memory Multi-Processor system
- Also known as Distributed Shared Memory, DSM, system
- Flexible, modular, and scalable architecture
- Memory is physically distributed & virtually shared.

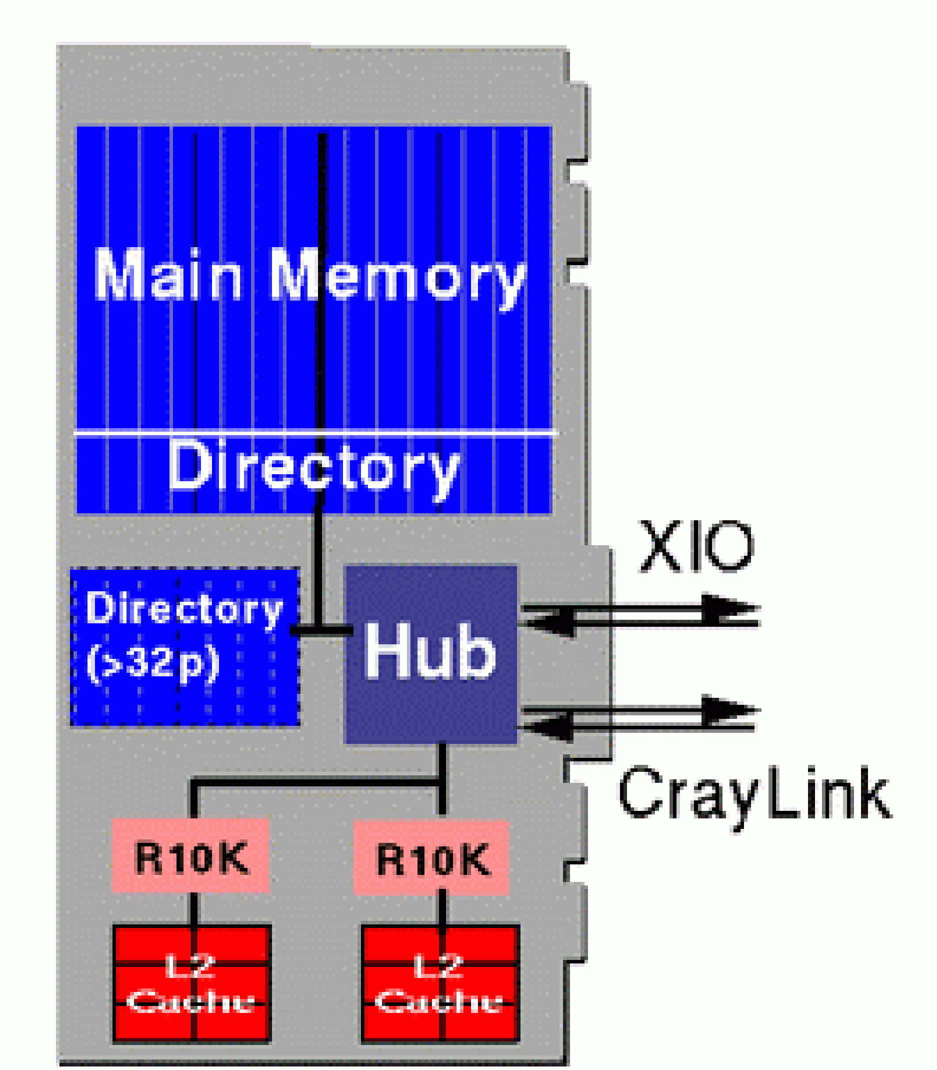
Overview (Continued)

- Offers ease of Shared-memory programming and scalability of Distributed Memory systems
- Scales in terms of the number of processors, memory size, I/O and memory bandwidth, and system interconnect bandwidth

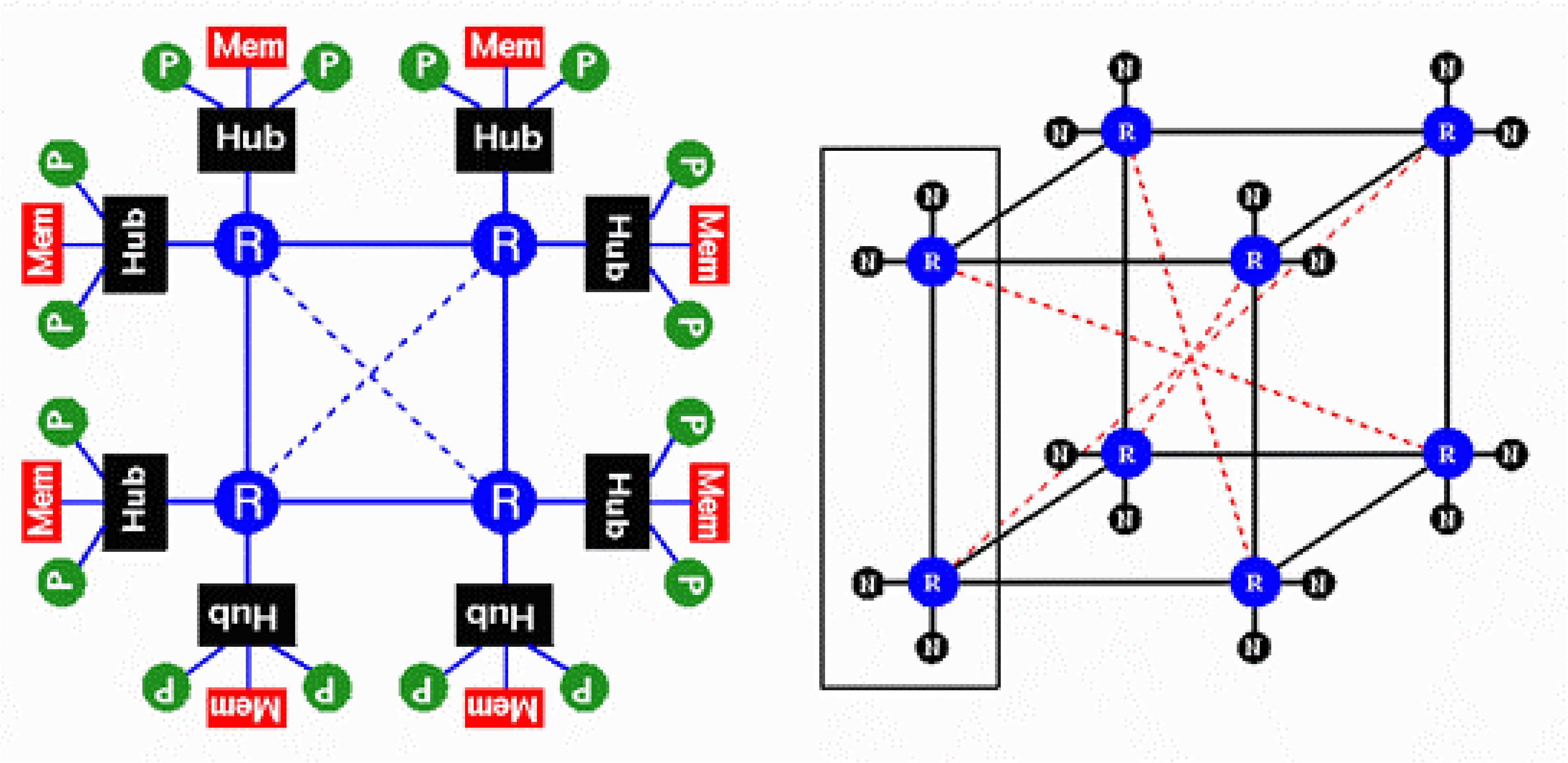
Origin2000 Building Blocks

- Built on the R10000 Processor
- The node card contains:
  - two R10000 CPUs
  - external caches
  - memory
  - HUB
  - I/O and interconnect interfaces
- High speed interconnect fabric - routers and proprietary links called CrayLinks

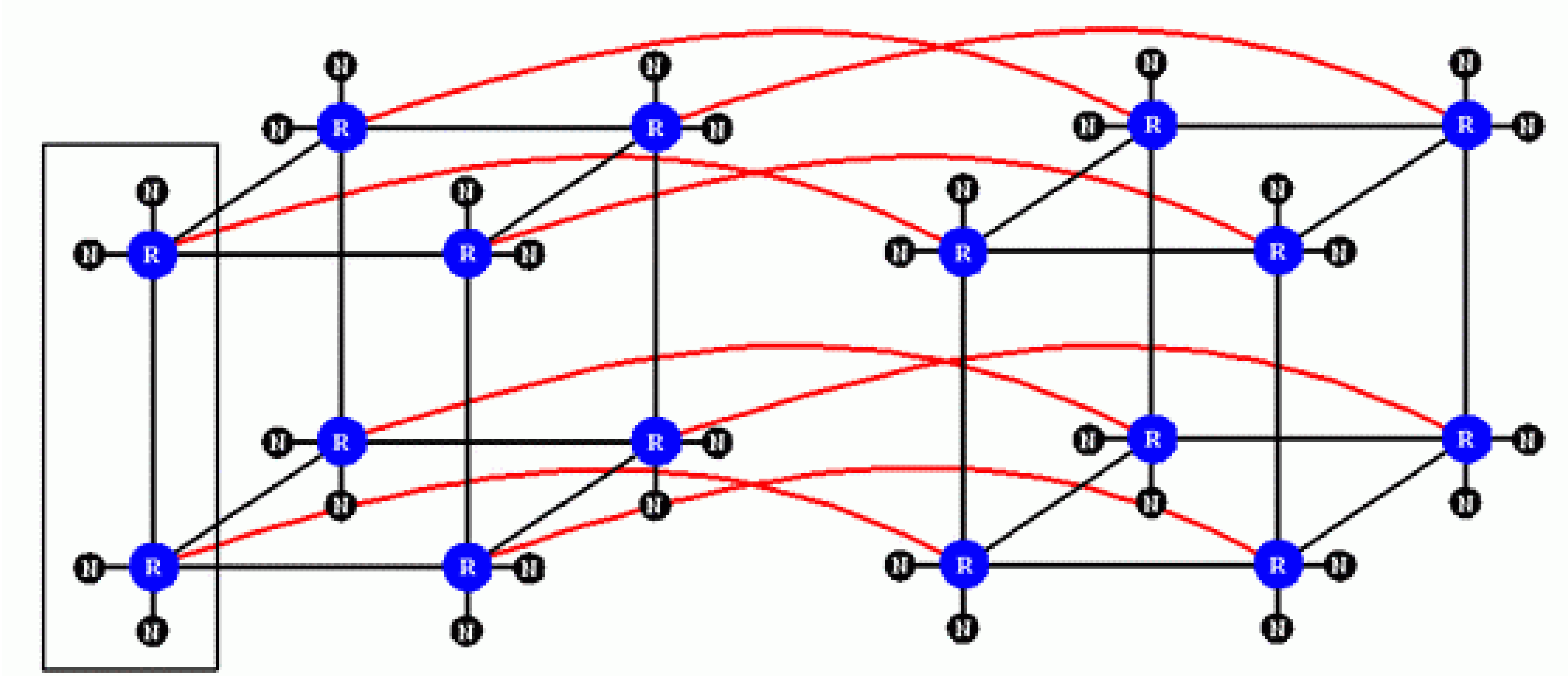
Origin2000 Node Board



16 and 32 Processor systems



64 Processor system



Directly connect two 32–node systems via Craylink cables using the one free link on each router

R10000 Processor

- **Runs @ 195 MHz or 250 MHz Clock speed**
- 64-bit processor
  - 64-bit registers
  - 64-bit integer and floating point ops. in HW
  - 64-bit virtual address space
  - 40-bit physical address space (1 TB)
- 5 independent fully pipelined functional units:
  - 2 FP units
  - 2 Integer units
  - 1 load/store unit

R10000 processor (continued)

- 4-way superscalar (up to 4 instr. issued/cycle)
  - up to 2 floating point instr. per cycle
  - up to 2 integer instr. per cycle
  - up to 1 load/store per cycle
- MIPS IV instruction set, binary compatible with earlier MIPS instruction sets
- Dynamic scheduling, Out-of-order instruction execution (when no instruction dependencies)
- Branch prediction

R10000 Processor (continued)

- Peak Performance (at 195 MHz / 250 MHz)
  - 1 mult-add/cycle (chained) OR 2 FP instructions
  - 390 / 500 MFLOPS: 2 floating point ops. per cycle
- MIPS rating
  - 4 instructions / cycle
  - 780 / 1000 MIPS

Cache on R10k Processor

- L1 Cache
  - 32KB floating point data
  - 32KB integer/instruction data
- L2 Cache
  - 4MB per processor
  - 2-way set associative; 2 banks each
  - cache line 128 bytes
  - cache clock rate
    - 2/3 that of CPU for 250 MHz CPU
    - same as that of CPU for 195 MHz CPU

Memory Subsystem

- Local memory on node card, shared by two CPUs
- Max. of 4 Gbytes of memory per node card
- 4-way interleaved (multiple memory accesses)
- ~ 6% of local node memory in < 32-processor configuration is used by the directory (for cache coherency)
- for systems > 32 processors, additional directory memory is needed (~15% of local memory)

Cache Coherency Procedure

- Associated with each cache-line size of memory are extra state presence bits which indicate which processors have a copy of that part of memory
- When a processor fetches a cache-lime form memory, it gets the data and the state presence bit for that processor is set
- To modify:
  - Gain exclusive ownership
  - Retrieves the state presence bits
  - Invalidate sent to all other processors
  - Others discard cached copies
  - Other processors get the fresh data from owners cache
- On write to main memory processor relinquishes ownership

HUB

- HUB is a crossbar switch on the node card
- Links : the two CPUs, local node memory, system interconnect (through a router), and

  I/O subsystem
- The two CPUs access local shared memory through HUB (similar to a bus-based system)
- Resolves memory addresses requests, and sends to local memory, or to remote memory through router

CrayLink Interconnect

- Router
  - on each node, connects the HUB to the Craylink interconnect system
  - 6-port switch
  - Determines most efficient connection to route a message
- Craylink Interconnect
  - Links two routers (or HUBs)
  - Bi-directional interconnect
  - 780 MB/s peak bandwidth in each direction
  - ~600 MB/s effective bandwidth for user data

Programming the Origin2000

- Supports SGI specific
  - Parallelization directives
  - Shared memory copies
  - References:
    - http://www.arc.unm.edu/Workshop/SMP/SMP_workshop/SGI_data_placement/SGI_data_place.html
    - http://www.arc.unm.edu/Workshop/SMP/SMP_workshop/SGI_directives/SGI_directives.html
  - Automatic parallelization
- Suggested use:
  - MPI
  - OpenMP